Annual Progress Report

Working Group:	***International Climate Network Working Group (ICNWG)***
Organization:	*Earth System Grid Federation (ESGF)*
Principle Investigator:	*Dean N. Williams, LLNL/AIMS*
Lead Engineer:	*Eli Dart, LBNL/ESnet*

January 2015

Introduction

In November 2013, the Enlighten Your Research-Global (EYR-Global) program awarded Dean N. Williams, project lead for the Earth System Grid Federation[1] (ESGF), and four other co-principle investigators the network resources and engineering skills requested in the EYR-Global'13 proposal "International Networking for Climate." Specifically, this project requested and was granted the resources for improving connectivity between five international ESGF sites for improved data replication performance. The participating ESGF sites included:

1. Analytics and Informatics Data Management Systems (AIMS), located at Lawrence Livermore National Laboratory (LLNL), US;
2. Centre for Environmental Data Archival (CEDA), located at Rutherford Appleton Laboratory (RAL), UK;
3. German Climate Computing Center (Deutsches Klimarechenzentrum GmbH, DKRZ), DE;
4. Royal Netherlands Meteorological Institute (KNMI), NL;
5. National Computational Infrastructure (NCI), located at the Australian National University, AU.

To help meet the needs of this proposal, the EYR-Global program appointed a lead network engineer to lead the performance improvement and engineering efforts. There after, the lead engineer, Eli Dart (ESnet), and proposal PI, Dean N. Williams, formed a working group within ESGF called the International Climate Network Working Group (ICNWG).

The ICNWG is currently supported in collaboration with an international group of national research and education networks (NRENs): AARNet (Australia), DFN (Germany), ESnet (United States), Janet (United Kingdom), and SURFnet (Netherlands), with additional resources and connections made via Internet2 (United States) and GEANT (the Pan-European research network).

However, there are many other organizations that are involved in helping this working group achieve its milestones as listed in Appendix B. Figure 1 shows the institutions and organizations supporting ICNWG and how they overlap. The working group also created a website[2] to make the ICNWG's progress transparent to all participants.

[1] ESGF is a coordinated multi-agency, international collaboration of institutions that continually develop, deploy, and maintain software needed to facilitate and empower the study of climate change. Through ESGF, users access, analyze, and visualize data using a globally federated collection of networks, computers, and software. See http://esgf.llnl.gov for more information.

[2] https://icnwg.llnl.gov

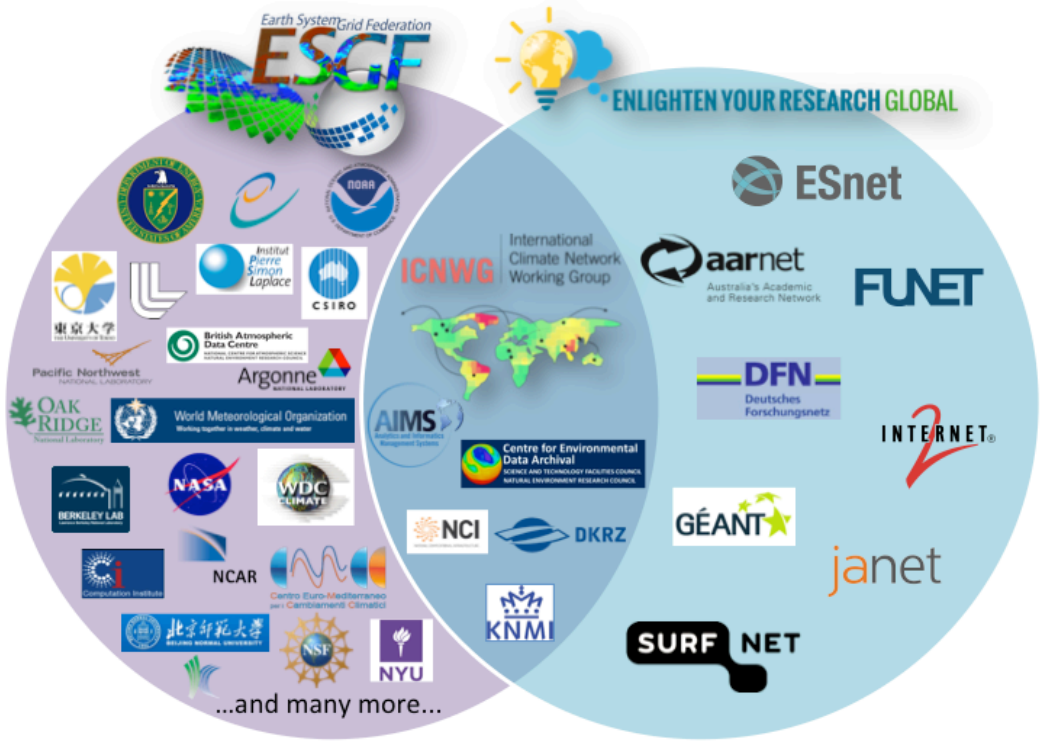


Figure 1 Diagram showing the overlap between the ESGF organization and the EYR-Global Program. Data centers noted in the crossover region are supported by the EYR-Global network organizations (on the far right). ESGF has many more institutions that participate in the organization, however, five data centers are pioneering how to upgrade or update their infrastructures to improve large-scale data replications.

As ESGF paves the way for a new era in climate system analysis and understanding, ICNWG will put into place a new set of networking best practices to effectively transport hundreds of petabytes of future large-scale climate data. The five data centers looking to achieve consistent disk-to-disk data transfer rates range from 4 to 8 gigabits per second (Gbps), or 0.5 to 1 gigabyte per second (GB/s). This can also be thought of as transferring approximately 1PB per month (4Gbps) and 1 PB per 14 days (8Gbps).

Enabled by the EYR-Global program, the network connections made and/or improved through the ICNWG will help climate and computational scientists manage and disseminate petabytes of modeling and observational data, which will traverse more than 13,000 miles of networks, spanning two oceans, and three continents.

Overview of Progress

When starting this project, data transfer performance was variable between all sites, between 10KB/sec and 40MB/sec. In many cases, the software used to transfer the data is wget or similar http-based tools. ICNWG aims to transition the collaboration to GridFTP for data replication, driven by Globus. The data set to be used for testing replication performance will be acquired within the month of February 2015 to test the disk-to-disk and memory-to-memory performance and provide a realistic approximation of production-level data replication performance.

In general, local network, server, and storage infrastructures must be upgraded to accommodate the data transfer performance required for this project. Many sites have made significant

progress in these areas and we have enumerated the details of their work in this report. (See Appendix A figure.)

In completing one of the first major milestones, four out of the five participating ICNWG sites have deployed perfSONAR nodes. perfSONAR measures the network performance capabilities at the end sites by using the tools bandwidth test controller (*BWCTL)* (for throughput testing, run every few hours) and One-Way Active Management Protocol (*OWAMP)* (low-bandwidth one way delay measurement and packet loss testing, running continuously). The results are then stored on a server, which can be viewed using a web interface, MaDDash (Monitoring and Debugging Dashboard), for easier performance troubleshooting and understanding (see Figures 2 and 3).

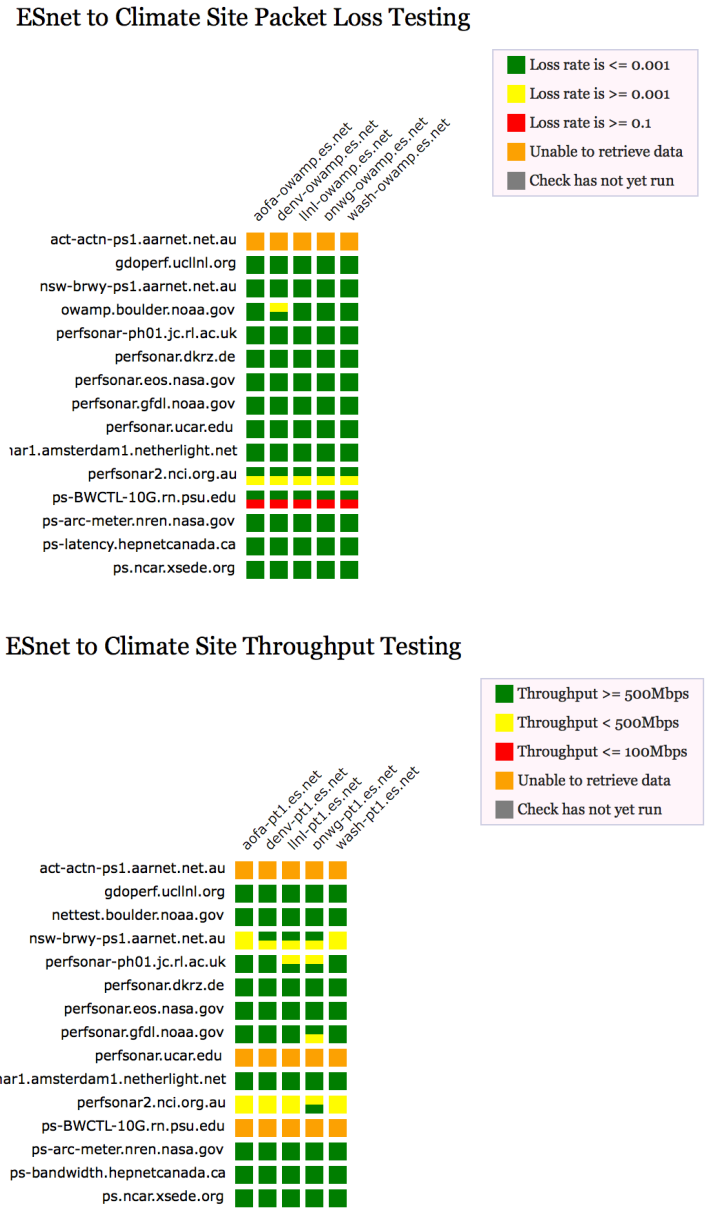


Figure 2. (Top) OWAMP dashboard showing packet loss between ESnet perfSONAR boxes and ICNWG participating organizations. (Bottom) BWCTL dashboard showing bandwidth tests between perfSONAR nodes (pt=performance testers). These are single-stream test results. When running tests four-way parallel, some

sites are meeting network speeds already. (Please note, network organizations and data centers have or already had perfSONAR deployed which makes testing network paths easier.)

In the next six months, the ICNWG will have its own server, icnwg.es.net, to accumulate data between all the perfSONAR nodes. The server is hosted and managed by ESnet. This will allow all nodes to test to each other in a mesh-like environment, which will show more realistic results for data replication that may not traverse ESnet, between Australia (NCI) and Germany (DKRZ), for example.

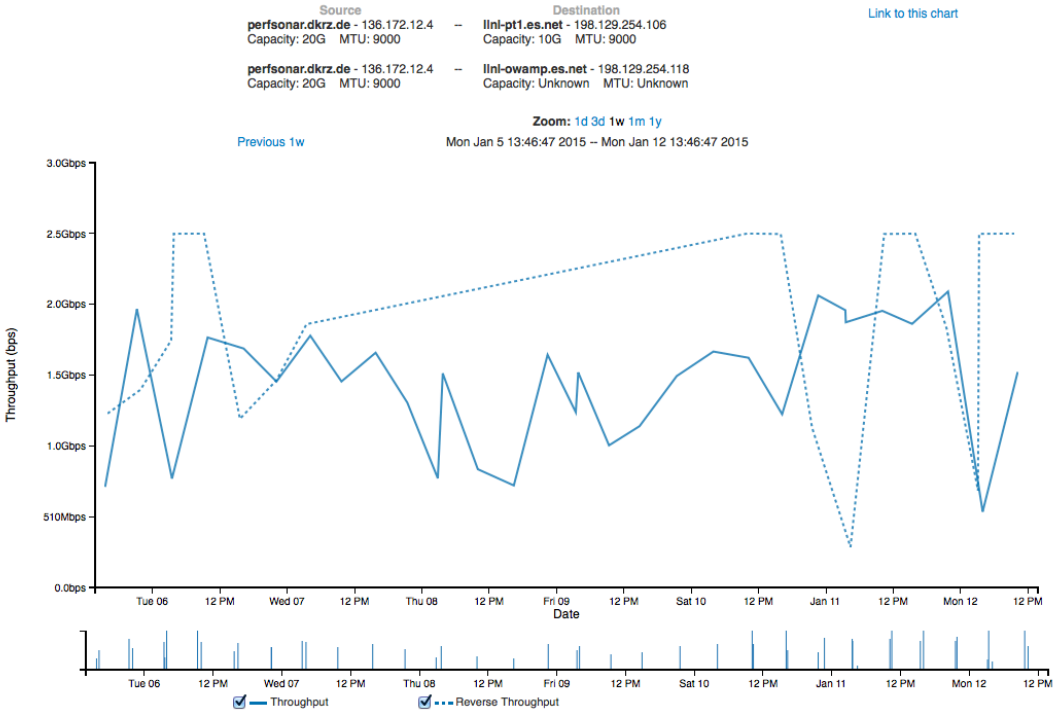


Figure 3. By clicking on a cross section of the dashboard grid, further information is available going back to one year. This example shows the BWCTL tests between DKRZ and the Lawrence Livermore National Laboratory (where AIMS is located) for a one-week time period (note the "Date" on the x-axis spans from Monday, January 5 to Monday, January 12, 2015). Notice bandwidth reaches an average of 2+Gbps for single-stream tests. By looking at the trends in the BWCTL and OWAMP graphs, engineers are able to narrow down issues in network performance due to failing hardware, asymmetric routing behavior, etc.

Specific Site Progress

Below we discuss and enumerate the progress and work at each site to obtain the goals stated in Appendix B.

AIMS, LLNL

The Livermore Computing (LC) team has performed work on the AIMS network at LLNL with contributions from the System Administration Group, Networking Team, and Advanced Technology Office (ATO).

Within the last year, the LLNL team completed a Technical Report, entitled, "LC Recommendations for Data Transfer Nodes," that summarizes the results of the team's extensive evaluation of different DTN configurations and also includes specific hardware recommendations. The LLNL team tested disk-to-disk data transfers with GridFTP on both virtual machines (VMs) and bare-metal servers, using various networking configurations on each. Each node type performed equally well, exceeding 800 MB/s with multiple streams. Due

to the simplicity of bare-metal server configuration and administration, LLNL recommends their use; consequently, LLNL's updated recommendations do not utilize virtual machines. Based on these findings, a hardware purchase went out in early January for the production DTN cluster, which will consist of four total nodes—three worker nodes and one management node. Delivery for the DTNs is expected by early March, with deployment of the DTNs completed one month later.

The current interim DTN node was configured to run Globus Connect (GC) and supports GC and GridFTP connections. It has actively supported numerous transfers, including large-scale transfers from the Australian site (NCI), summing to 183 TB so far. In addition, rigorous security measures including switch ACLs and firewall rules were put into place to protect the resource.

Another recent accomplishment was participation by LC staff in the Science DMZ testing at the Supercomputing Conference (SC14) in November 2014. This was an effort involving Cisco, SCinet, and LLNL to create a temporary Science DMZ utilizing 40GbE and 100GbE network architectures. The goal was to determine the best hardware and tuning to achieve the best rates on high bandwidth networks. Eventually LLNL found a hardware combination that achieved the absolute theoretical maximum throughput from a 40.0Gbps network interface card (NIC). The team reached this performance level using bwctl iperf, setting a window size of 128MB and running four-way parallel streams ("-P4 -w 128M").

The connection from the LLNL Science DMZ to ESnet is in the process of being upgraded to 100Gbps from 10Gbps. LLNL expects the upgraded infrastructure to be completed and put into production in early 2015. The upgrade will significantly increase the network capacity available for data transfers between the AIMS data center and the other climate data centers.

The required resources for AIMS to participate in the project include approximately three person-months completing initial benchmarking, hardware acquisition, installation, requirements gathering, research, and configuration.

One challenge faced at LLNL is the lack of a Globus Connect instance that includes the ESGF Authz (authorization) module. For now, user administration is a manual process, but once Globus provides an updated product, all ESGF users will be able to access the LLNL DTN system via their ESGF OpenID accounts and MyProxy.

In the next six months, AIMS will deploy a production DTN cluster and updated GC instance. Additional tuning and configuration of the DTN cluster will be performed to ensure current and future goals are met.

The design of the data storage system may need to be altered in order to meet the longer-term stretch bandwidth goals. Incorporating a parallel file system may be required if AIMS staff find the Network File System (NFS) performance to be inadequate.

CEDA, RAL

Work undertaken so far at CEDA (or on behalf of CEDA by STFC's Scientific Computing Department) has been done in parallel with a major infrastructure project, JASMIN (www.jasmin.ac.uk), which has deployed petascale high-performance storage, virtualization and cloud computing capabilities for environmental science, along with site networking changes. A key part of this has been setting up a "Science DMZ" as part of the JASMIN network: a dedicated "friction-free" zone outside the site firewall. Within this zone, staff members have deployed (on

10G-connected physical hosts):

- A perfSONAR node,
- A high-performance FTP server for user access to CEDA archives, and
- A high-performance data transfer server for user (RW) access to collaborative workspace areas.

With help from ESnet engineers, the perfSONAR host (perfsonar-ph01.jc.rl.ac.uk) is now operating with both packet loss[3] and throughput[4] tests. Recent tests have achieved up to 3.8Gbps in throughput tests to some ESnet hosts with 0% packet loss.

Alongside the Science DMZ, in the "standard" infrastructure, are the original service hosts (which the high-performance hosts may replace at some stage). These are hosted as virtual machines within a 10G-connected VMWare virtualization environment. An additional perfSONAR node (a virtual machine) has been deployed alongside these "standard" hosts as representative of that environment.

Further work has concentrated on deploying and testing GridFTP on a test server in the standard infrastructure. Configuration and testing is still ongoing, but once complete will be followed by GridFTP deployment on server(s) in the Science DMZ.

Parts of the project specifically related to ICNWG goals amounts to approximately two staff weeks of system support, plus additional management and other activities. System support effort concentrated on deployment of physical and virtual machines, liaisons with site networking to affect network changes, firewall configuration and interaction with ESnet engineer support. The main challenges for CEDA, to date, have been balancing priorities for staff time against the larger task of constructing and integrating the JASMIN infrastructure, and understanding the perfSONAR configuration (how to set up tests with firewall configurations).

In the next six months, CEDA within the JASMIN infrastructure will

- Monitor and study perfSONAR results to ensure best performance is being achieved
- Deploy GridFTP on existing high-performance data transfer node in Science DMZ
- Set up of data transfer tests with other data centers and test disk-to-disk performance
- Set up of ESGF data node webserver in Science DMZ
- Set up of additional data transfer node for arrivals in Science DMZ (for replication use case, with twin for striped GridFTP use)
- Investigation use of hypervisor in Science DMZ for hosting additional nodes (limited space for physical servers). Includes setup of dedicated hypervisor
- Set up additional perfSONAR tests with other partner sites in UK

Setting up a Science DMZ has already *shown major (10x) performance benefits* in throughput for end user science data transfers (exceeding end-user expectations in capability for data moving).

Understanding long-path network transfers has increased among system support and data center staff, who are now better able to advise users on appropriate transfer methods for

[3] This test is a one-way ping test: the One-Way Active Management Protocol (OWAMP). More information can be found at http://software.internet2.edu/owamp/.

[4] This test is a bandwidth test controller (BWCTL). More information can be found at http://software.internet2.edu/bwctl/.

particular scenarios.

And some foreseeable challenges looking forward include:

- Expansion of Science DMZ to additional physical hosts may be limited by physical space, hence investigating use of a hypervisor for Science DMZ VMs.
- Set up of additional server for striped GridFTP will likely require additional testing and troubleshooting.
- Contention with a growing community of bandwidth-hungry science data end users (demand growing to meet capacity) is also an ongoing challenge.

DKRZ

Motivated by feedback from CMIP5, the ESGF data federation is emphasizing network connectivity between key ESGF data nodes at DKRZ, CEDA, and AIMS. DKRZ operates together with AIMS and CEDA as the core data archive for the CMIP5 data federation. The data scale for such an archive ranges from several terabytes to petabytes. This replication is also required to be fast and secure.

DKRZ staff started this project by creating a clear picture of the current network topology including hops and bandwidths between sites. Second, they documented and gathered metrics about the network speeds and data transfer speeds. After identification of bottlenecks and adaptation of transfer parameters (server configurations, etc.) and routing paths, the data transfers between DKRZ and AIMS and between DKRZ and CEDA could be increased to considerably more than 1 Gbps, with peak rates above 3 Gbps.[5] DKRZ is connected to their national research network with a cumulative 7Gbps link. Network tests with the Australian partner, NCI, have not yet been successful.

In the next few months, network bandwidth rates should be stabilized and increased in the direction of 10 Gbps, providing the nominal bandwidth is sufficient. In addition, data transfers to and from Australia will be integrated in the tests. In conclusion, the end-to-end data transfer from one ESGF data node to another ESGF data node should be tackled with respect to the ESGF CMIP5 data replication use case.

NCI, ANU

NCI has relocated all ESGF data to the NCI global filesystem as of December 2013. The data at NCI was migrated to a new large Lustre filesystem, which will now serve all ESGF data within the region.

There was a delay in finalizing these nodes in NCI's virtual environment due to the ESGF software stack that had a flow on effect to how these core nodes could be deployed to the NCI cloud environment. The install process was resolved by late August 2014. Data transfers are currently taking place on generic data transfer nodes of NCI (r-dm5 and r-dm6). Direct data transfers using these nodes commenced in late July 2014 using direct logins to the other major nodes filesystems.

Filesystem testing was finished by May 2014 for the 10G data servers. There are several stages of testing and tuning of the data servers. The first phase of testing achieved 742MB/sec from `/g/data1` for memory to disk using a natively mounted node:

[5] More information about DKRZ test results can be found at https://perfsonar.dkrz.de/serviceTest/index.cgi?eventType=bwctl. Lower memory limits were active between June 19 and July 1.

```
% fs setstripe -s 1M -c 10 test/
% cd test/
% ~/bbcp/bbcp -t 30 -F -P 2 /dev/zero
localhost:/g/data1/ua6/sync/test/test.out
bbcp: Creating /g/data1/ua6/sync/test/test.out
bbcp: 140805 16:53:14  0% done; 742.4 MB/s
```

These results exceed the performance requirements for the May milestone. Currently the NCI is obtaining a maximum of 962 MB/sec for synthetic disk benchmarks run on the DTN node onto the NCI's global Lustre filesystem. Further investigation is ongoing to ensure these number are sustained during disk-to-disk WAN transfers.

NCI has deployed stand-alone data analysis nodes and a user environment in the NCI cloud. The first stage was completed in October 2013, and a second and production-ready phase was released by August 8, 2014. The data from CMIP5 is directly available for interactive analysis. The VisTrails environment (also being deployed within the UV-CDAT) has been made available and user workflows captured. A future phase will include integration with NCI data provenance capture system that will capture data replication from ESGF nodes, and their usage in the NCI cloud.

Major work for data transfers via the network did not commence on time due to a staff turnover at NCI and the later than expected filesystem delivery. NCI is connected to the Australian national research and education network, AARNet, via a 10Gbps link, which is then connected to a 100Gbps backbone network and to a dual trans-Pacific 40Gbps link both operated by AARNet.

Testing of network transfers took place in July 2014. Network transfer tests are plotted below, and indicate the average data transfer rate over the network (no disk) between NCI and a well-connected node at Lawrence Berkeley National Laboratory (see graph of metrics below). This indicates that it is possible to get a line rate of 1GB/s (8Gbps) between Australia and the US. Initial data replication has commenced from DKRZ, however, transfer tests to DKRZ have shown the network transfer performance to be very low—on the order of 80 Mbps. It is unclear if this is due to a local DKRZ issue or if it is due to a wide-area network issue between Australia and DRKZ (via trans-Pacific links, R&E networks in the US, trans-Atlantic links, or in the European R&E network domain, etc).

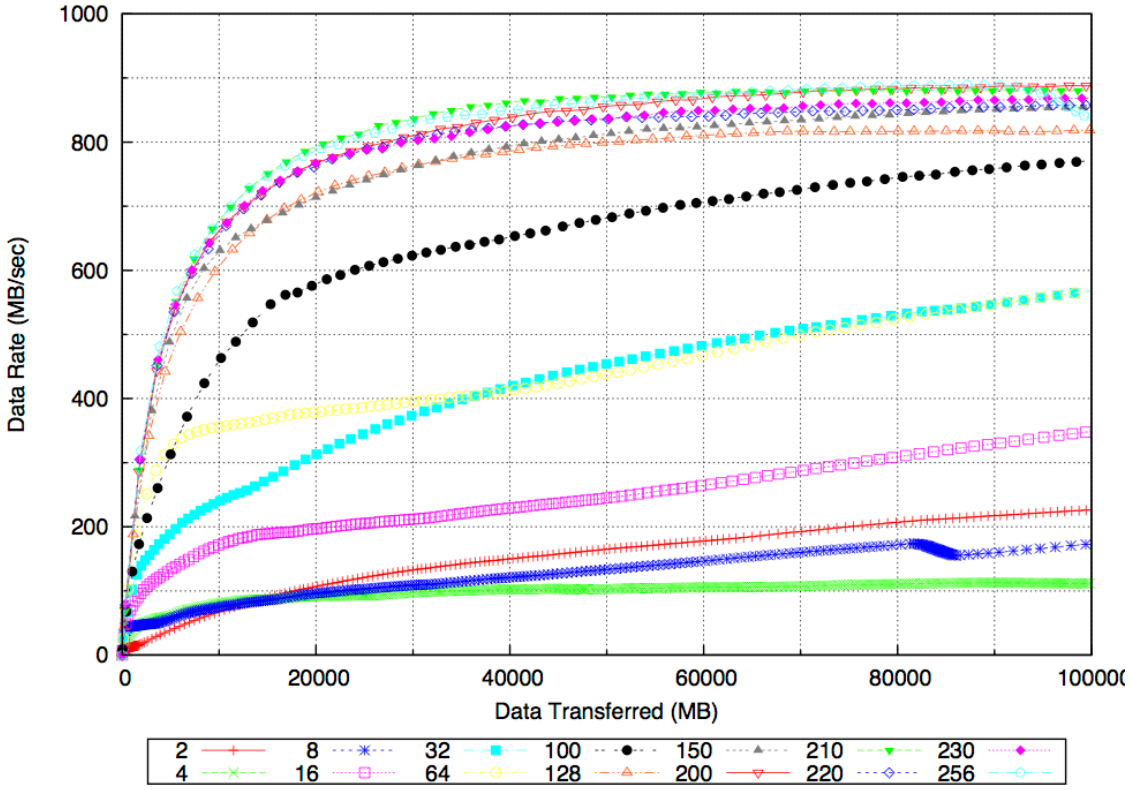


Figure 3 - Graph shows the data rate vs. the volume of data transferred. Different lines in the graph represent how many data streams were required to obtain the given performance. The results of the graph indicate that it is possible to get a line-rate of 1GB/s (8Gbps) between Australia and the United States, however, it requires configuring transfers to run more than 100 parallel streams. Discussions with ESnet have shown such large number of parallel streams indicate bottlenecks in the node and network configuration.

On July 24, 2014, initial statistics from using globus-url-copy using ESnet hosts yielded:

```
Memory to Memory using 100 TCP streams - lbl—diskpt1 to r-
dm6.nci.org.au:
31413501952 bytes       576.12 MB/sec avg       674.01 MB/sec inst

Memory to Disk using 100 TCP streams - lbl—diskpt1 to r-dm6 (/g/data1):
107373395968 bytes 278.26 MB/sec avg 362.58 MB/sec inst.

Memory to Disk using 100 TCP streams - lbl—diskpt1 to r-dm6 (/short):
107359502336 bytes 270.86 MB/sec avg 297.49 MB/sec inst

Disk to Memory using 100 TCP streams - lbl—diskpt1(/data1) to r-dm6
(/dev/null):
104856289280 bytes       364.96 MB/sec avg       722.96 MB/sec inst

Disk to Memory using 100 TCP streams - lbl—diskpt1(/data2)  to r-dm6
(/dev/null):
104261222400 bytes       350.11 MB/sec avg       343.87 MB/sec inst

Disk to Disk using 100 TCP streams - lbl—diskpt1 to r-dm6 (/short):
104802025472 bytes       234.62 MB/sec avg       184.00 MB/sec inst

Disk to Disk using 100 TCP streams - lbl—diskpt1 to r-dm6 (/g/data1):
104809365504 bytes       137.66 MB/sec avg       203.95 MB/sec inst
```

Discussions with ESnet have shown such large number of parallel streams indicate bottlenecks in the node and network configuration.

A number of perfSONAR boxes have been installed within AARNet, which are helping to troubleshoot local-area issues vs. wide-area network issues. The main test instance NCI is using is nsw-brwy-ps1.aarnet.net.au. NCI's local perfSONAR deployment is located within the NCI data center (perfsonar2.nci.org.au), however, staff are looking to deploy another physical host near NCI's DTN to help troubleshoot bottlenecks between their local network and the wide-area networks. AARNet's and NCI's perfSONAR installations are producing results on the ICNWG Dashboard at https://icnwg.llnl.gov. The results indicate that tests from AARNet to ESnet are quite variable with a maximum throughput of 1.9Gbps, a mean of 781 Mbps and a minimum of 162Mbps. The reverse direction is more constant at approximately 2.5Gbps. This indicates a performance asymmetry in the network, which appears to be caused by packet loss in the Australia to United States direction. An investigation of the network path is ongoing.

Within the next six months, installation of the latest ESGF node software (version 1.8) will be taking place.

Additional activities undertaken at NCI were:

- *An attempt at bulk replication of CMIP5 data from CEDA, DKRZ using bbcp*. Initial attempts were unsuccessful using bbcp and rsync owing to poor throughput. CEDA will look at setting up a dedicated very-wide-area tuned transfer host, as their current hosts are tuned for regional European transfers. With the DKRZ node, outbound flows were problematic and further investigation is required.

- *Bulk replication of CMIP5 data from LLNL using GridFTP*. Around November 25, 2014, NCI commenced bulk replication of CMIP5 data holdings from LLNL. Progress has been hampered at the NCI end due to the absence of a GridFTP server. A workaround is currently in place (discussed below).

- *Local NCI systems evaluations on existing computing resources, data transfer nodes, perfSONAR hosts, Lustre file system, GridFTP installations, etc.*

  Data mover nodes (r-dm nodes) of the Raijin supercomputer were used to facilitate initial GridFTP replication of LLNL's data. A significant body of work revealed TCP tuning, Lustre tuning and NIC tuning were required. As Raijin's r-dm nodes were not designed as a DTN, they are being used in a stop-gap manner until a DTN-based GridFTP server is setup. A copy of Globus Connect personal edition is currently running on r-dm6 and is used for a Globus-mediated transfer from the LLNL node to NCI. To date, this was used to transfer 152TB of data.

  Two dedicated DTN nodes and two perfSONAR nodes were provided to NCI, as part of a national academic networking project (DaShNet). One of the DTN nodes is being configured and benchmarked to run a GridFTP server. AARNet staff will manage and run two perfSONAR nodes at NCI. These nodes reside within the NCI local network and once commissioned, will provide metrics on the network throughput and loss. NCI is still waiting on an estimated deployment date for these nodes.

  Once single-stream performance and filesystem issues have been ironed out, work will commence to complete the NCI's GridFTP server install with Globus.

Work is being done to provide base-line filesystem performance statistics as seen from the DTN node. So far, memory-to-memory performance improved once shutting down unnecessary OS daemons. The improvements are shown in Figures 4 and 5 where NCI's percentage of bandwidth utilization performance went from a frequent saw-tooth pattern to a less frequent pattern. The second plot shows the saw-toothed pattern after changes were applied to the DTN node, showing a less frequent oscillation band (760MB/sec to 822MB/sec). Further work is under way to improve the interrupt request (IRQ) affinity and numa-based thread binding.

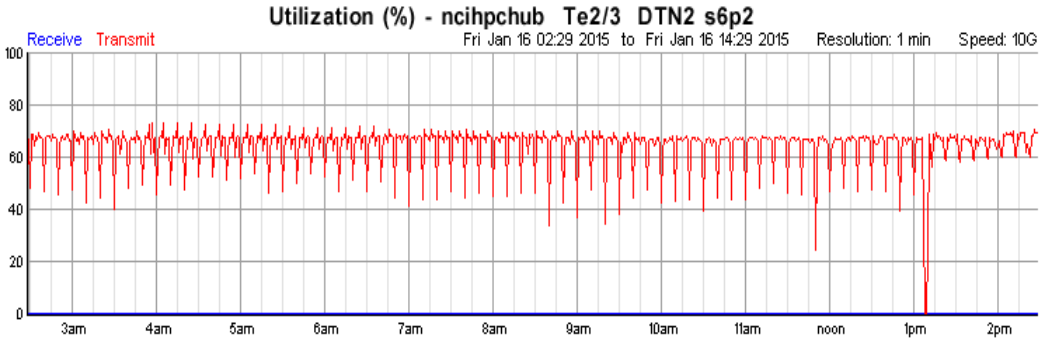


Figure 4 - This plot shows utilization oscillating between 70% and 45%, where the x-axis is time (January 16, 3am to 2:30pm); the y-axis is percent utilization of the DTN's 10Gbps uplink.

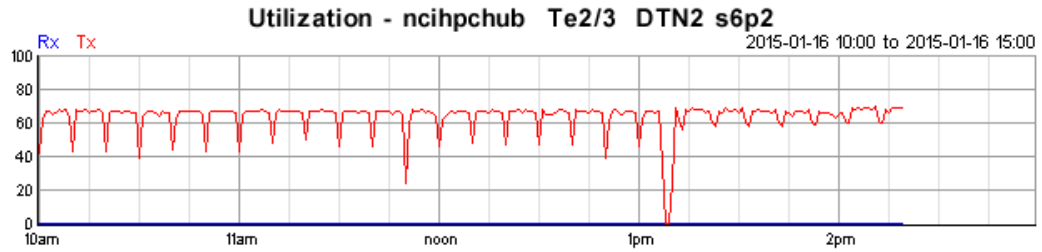


Figure 5 - The plot shows the saw-toothed pattern after changes were applied to the GridFTP server, showing a tighter more consistent transmission: 760MB/sec to 822MB/sec.

NCI staff also examined the DTN's PCI device placement. This showed that an Infiniband (IB) host channel adapter (HCA) used to access the storage infrastructure was attached via PCIe to the second CPU socket (see Figure 6), where as the Ethernet NIC involved in WAN transfers was attached to the first CPU socket. This arrangement leads to CPU cache-coherency traffic traversing the QuickPath Interconnect (QPI) link between sockets, as the storage HCA is not on the same PCIe bridge as the network card. A CPU thread reading data from the Ethernet NIC on the first CPU, when writing data to the IB HCA must necessitate serialization-deserialization to data using the QPI links between the two sockets which is a slow-path operation. NCI is working towards remedying this.

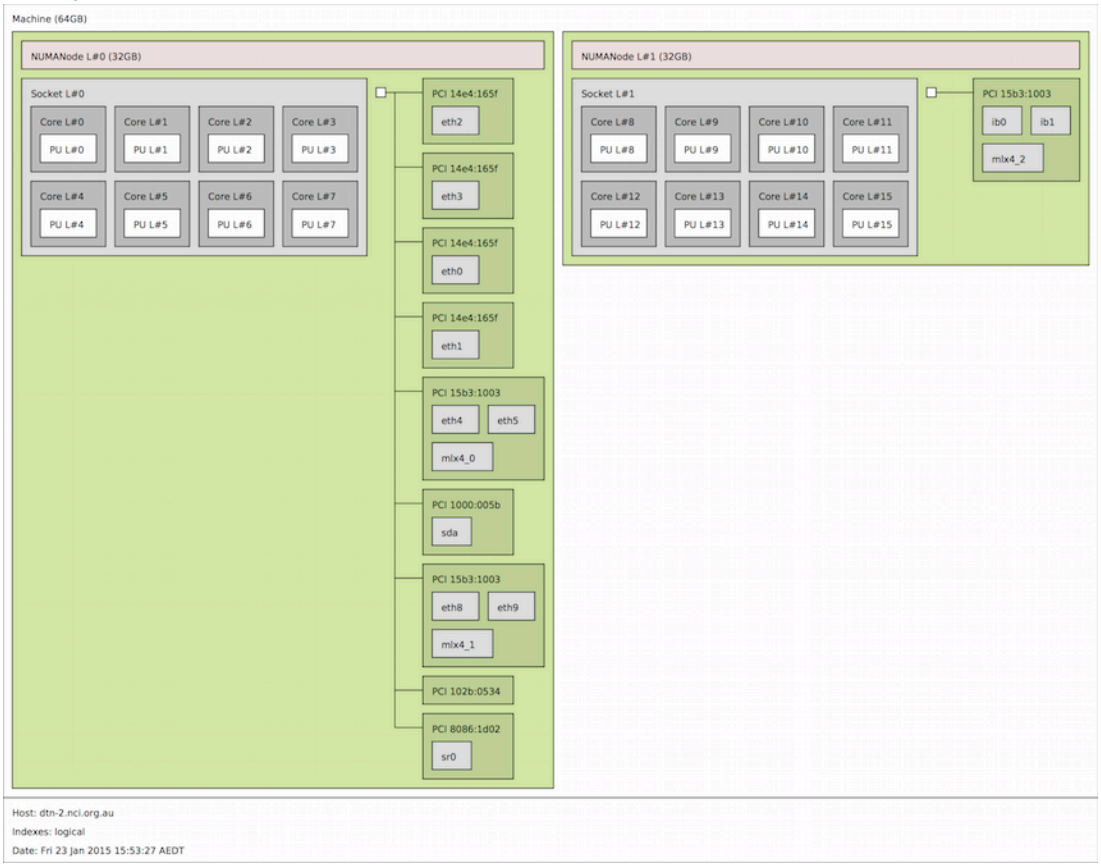


Figure 6 - This arrangement leads to CPU cache-coherency traffic traversing the QPI link between sockets, as the storage HCA is not on the same PCIe bridge as the network card: i.e, a CPU thread reading data from the Ethernet NIC on the first CPU, when writing data to the IB HCA must necessitate serialization-deserialization to data using the QPI links between the two sockets which is a slow-path operation.

In August 2014, NCI's Internet link was upgraded by ANU Netcomms and AARNet to the new AARNet4 network. The previous primary AARNet3 connection had a practical limitation on throughput of around 5Gbps on a link speed of 10Gbps. The new AARNet4 connection allows full speed across the link.

The first phase of this work involved migrating both the NCI edge from a Cisco 6509 to a new Cisco 6807 (deployed directly in the NCI DC) and the AARNet CPE from a Cisco 7604 CPE to a Juniper EX4550 NTU.

The second phase of this work (currently in progress) will upgrade the backup connection to the same standard as the primary Internet. These will allow for complete build and fibre diversity as well as allow for load balancing of Internet traffic under certain circumstances.

Significant planning and scheduling was needed due to the multiple NCI partners needing minimal downtime for access to NCI resources (supercomputing, cloud and storage). The next few months will concentrate on bringing up NCI's secondary link, allowing load balancing traffic and additional work on performance and security at the edge. This was completed Q1, 2015 and NCI will replace its secondary WAN edge from the existing Cisco 6503 to a new platform for service delivery over the next 5 years (to be delivered by Q4, 2015). We also need to migrate our existing direct links into NCI onto the new edge and we may bring on further direct links as AARNet4 is further rolled out (and allows for more cost effective point to point high speed

network connections).

Up to this point, some challenges at NCI have been:

- The ESGF node software and complications to install and federate in the proposed configurations,
- Discovering and progressing network performance issues,
- perfSONAR installation details are not cloud-ready,
- Local storage and cloud connectivity issues,
- Getting access to engineering resources on the network side of things – within NCI, with ANU Netcomms and from AARNet,
- Access to the remote nodes.

Looking forward, there are undefined challenges for NCI in using the ESGF nodes and GridFTP servers with the appropriate striping for production-level at this stage. The replication software and procedures are not yet fully known either and will take some more work.

Conclusion

Overall the ICNWG project is a huge undertaking, which requires coordination between different countries with different policies, and traversing multiple network domains. To improve network connections from end-to-end, summing to thousands of miles of fiber, and tens of network domains, this group will continue to work to support reliable, long-scale data transfers across the international domains. The ICNWG will continue to test and troubleshoot the network paths to each site, with ESnet's help, for the duration of the project (up until 2016).

The first goal for ICNWG was to achieve 500MB/sec (4Gbps, approximately 1PB per month) of disk-to-disk throughput during the replication of a multi-terabyte data set, using production infrastructure. This still has to be completed between all the sites but there is sure progress that the working group sites will reach this goal within the next 6 months. In 2015, this year, the ICNWG will increase that performance by a factor of two (to 1GB/sec, 8Gbps, 1PB per 14 days) between nodes. Achieving this capability on production systems will help prepare the global climate science infrastructure for the demands of CMIP6, and set the stage for continued scientific productivity in the critically-important area of climate science. As a stretch goal, if possible, the group wishes to try for a second doubling of performance of 2GB/sec (16Gbps, more than 1PB per week) between centers in 2016.

Appendix A

perfSONAR boxes have been deployed at most of the participating ICNWG sites. The below diagram shows boxes deployed near, and in the path of, the ICNWG sites, which can be used for testing.

Methods for testing between sites are performed using

```
bwctl [-c] [-s] hostname -t 30 -f m -i 2 -x -P 4
```

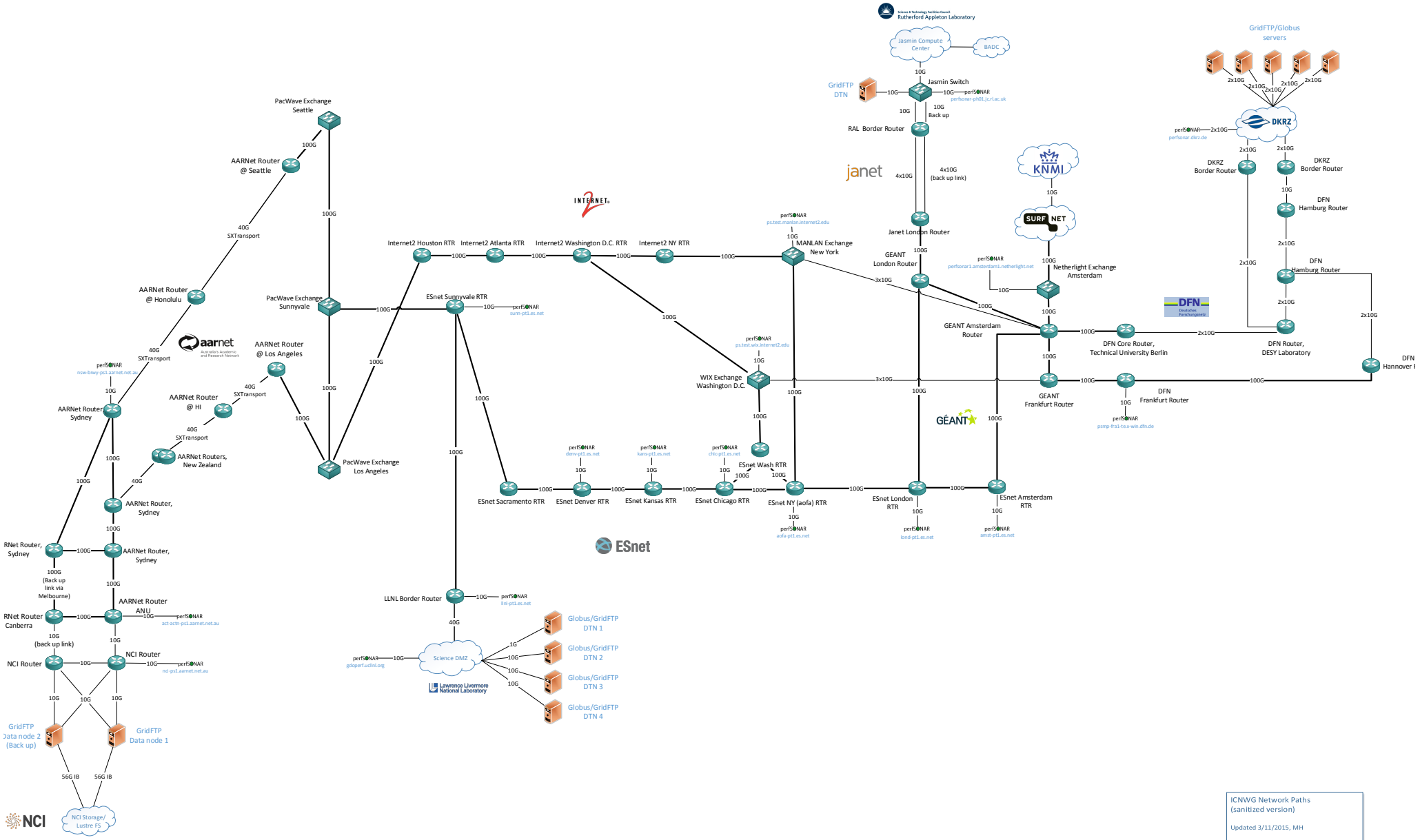
or by running a third-party test

```
bwctl -c hostnameA -s hostnameB -i 2-t 30 -f m -P 4
```

(in four-way parallel, -P 4). Loss and latency are tested using

```
owping -c 10000 -i .01 hostname.
```

Automated perfSONAR tests are run every few hours (usually for the duration of 20–30 seconds) 24/7 between sites to measure throughput and packet loss between sites. Test results are visible via the perfSONAR Climate Mesh Dashboard on icnwg.llnl.gov. Please note that automated tests in the mesh are single-stream since they are used to identify or indicate where potential problems or failures may be at various points in the data's route. Data replication transfers will be performed using at least four-way parallel streams, yielding higher throughput rates by a factor of 4.


ICNWG
International Climate Network Working Group
ESGF
Earth System Grid Federation
Rutherford Appleton Laboratory
Jasmin Compute Center
BADC
GridFTP DTN
Jasmin Switch
perfsonar-ph01.jcr1.ac.uk
Back up
RAL Border Router
janet
4x10G
4x10G (back up link)
Janet London Router
GridFTP/Globus servers
DKRZ
perfsonar.dkrz.de
DKRZ Border Router
DFN Hamburg Router
KNMI
SURFNET
Netherlight Exchange Amsterdam
perfsonar1.amsterdam1.netherlight.net
GEANT London Router
GEANT Amsterdam Router
GEANT Frankfurt Router
DFN Core Router, Technical University Berlin
DFN Router, DESY Laboratory
DFN Frankfurt Router
psmp-fra1-tix-win.dfn.de
DFN Hannover
GÉANT
INTERNET2
Internet2 Houston RTR
Internet2 Atlanta RTR
Internet2 Washington D.C. RTR
Internet2 NY RTR
MANLAN Exchange New York
ps.test.manlan.internet2.edu
WIX Exchange Washington D.C.
ps.test.wix.internet2.edu
ESnet Wash RTR
ESnet Sunnyvale RTR
sunn-pt1.es.net
ESnet Sacramento RTR
ESnet Denver RTR
denv-pt1.es.net
ESnet Kansas RTR
kans-pt1.es.net
ESnet Chicago RTR
chic-pt1.es.net
ESnet NY (aofa) RTR
aofa-pt1.es.net
ESnet London RTR
lond-pt1.es.net
ESnet Amsterdam RTR
amst-pt1.es.net
ESnet
PacWave Exchange Seattle
AARNet Router @ Seattle
AARNet Router @ Honolulu
AARNet Router @ Los Angeles
AARNet Router @ HI
AARNet Routers, New Zealand
PacWave Exchange Sunnyvale
PacWave Exchange Los Angeles
aarnet
nsw-bnwy-ps1.aarnet.net.au
AARNet Router Sydney
AARNet Router, Sydney
RNet Router, Sydney
(Back up link via Melbourne)
AARNet Router ANU
act-actn-ps1.aarnet.net.au
RNet Router Canberra
(back up link)
NCI Router
nci-ps1.aarnet.net.au
GridFTP Data node 2 (Back up)
GridFTP Data node 1
56G IB
NCI
NCI Storage/ Lustre FS
LLNL Border Router
llnl-pt1.es.net
Science DMZ
gridperf.ucllnl.org
Lawrence Livermore National Laboratory
Globus/GridFTP DTN 1
Globus/GridFTP DTN 2
Globus/GridFTP DTN 3
Globus/GridFTP DTN 4
ICNWG Network Paths (sanitized version)
Updated 3/11/2015, MH
Filename: Climate-ICNWG-full-clean.v6.vsd

Appendix B

Below is the original timeline for achieving the ICNWG's goals for each site at AIMS, CEDA, DKRZ, NCI, and KNMI. Red notes the revised timeframe for achieving the working group's the goals.

March 2014–September 2014
• Deploy 10G perfSONAR test server
• Deploy 10G data server
• Set up perfSONAR tests

~~**May 2014**~~ **July–December 2014**
• Filesystem tests for 10G data servers – target 500MB/sec
• Achieve 500MB/sec (4Gbps) network test throughput between perfSONAR test servers

~~**August 2014**~~ **January–February 2015**
• 500MB/sec (4Gbps) disk to disk transfers between data servers

~~**September 2014**~~
~~• Extra time for resolving issues~~

~~**November 2014**~~ **February–March 2015**
• Deploy second 10G data server
• Configure second 10G data server for striped Globus/GridFTP transfers with first 10G data server

March 2015
• Test striped Globus/GridFTP transfers with one other center

June 2015
• Test striped Globus/GridFTP transfers with all centers

August 2015
• Demonstrate 1GB/sec (8Gbps) transfers between all centers

Remainder of 2015
• Extra time for schedule slip
• Prep for 2016 stretch goals

June 2016
• Demonstrate 2GB/sec between all centers that are capable (stretch goal)

Appendix C

List of collaborators for the *ICNWG*.

AIMS, LLNL

- Dean Williams, LLNL's AIMS Program Leader, ESGF Project Lead, UV-CDAT Project Lead
- Robin Goldstone, Network
- Jeff Long, Security
- Cameron Harr

CEDA, RAL

- Matt Pritchard, CEDA Operations Manager
- Philip Garrad, Network
- Robin Blowfield, Security
- Cristina del Cano Novales, Security

DKRZ

- Michael Lautenschlager, DKRZ Department Head for Data Management
- Thomas Kaule, Network
- Gerald Vogt, Security

KNMI

- Wim Som de Cerff, Head of Global Climate Division
- Jeroen van der Reijden, Network
- Anita Hoegee-Wehmann, Security

NCI, ANU

- Ben Evans, Associate Director (Research) NCI at Australian National University
- Darren Coleman, Network
- Allan Williams, Security
- Joseph Antony, Data-intensive Projects
- Jason Andrade, Storage and Infrastructure
- Andrew Howard, Systems and Infrastructure

Globus

- Rachana Ananthakrishnan
- Lukasz Lacinski

List of network organization collaborators.

AARnet

- Guido Aben, Director of eResearch
- Warrick Mitchell, Network Engineer

DFN

- Stefan Piger (Jacob Tendel/Robert Stoy)

ESnet

- Eli Dart, Science Engagement Network Engineer
- Mary Hester, Science Engagement Coordinator, EYR-Global Program Team
- Jason Zurawski, Science Engagement Network Engineer

Janet

- David Salmon, Research Support Team Manager

SURFnet

- Sylvia Kuijpers, Community Support, EYR-Global Program Lead